Research Statement

As statistical models grow in size and complexity to serve modern scientific datasets and questions, fundamental data science tasks become more computationally onerous due to the need of many classical procedures to evaluate or fit a model multiple times. My research uses *sensitivity analysis* to provide fast, accurate approximations to such fundamental data science tasks, often exhibiting good accuracy and orders-of-magnitude speedup over classical methods.

Consider, as motivating examples, the following ubiquitous data science tasks.

- Cross validation (CV) is a fundamental tool in machine learning to evaluate model predictive performance and tune hyperparameters, but requires fitting a model multiple times with different data subsets left out.
- Prior specification is a necessary step in Bayesian statistics, a statistical paradigm that provides interpretable, coherent uncertainty quantification for scientific questions. But Bayesian inference can be sensitive to the prior specification, and evaluating the model for multiple plausible prior choices can be computationally prohibitive.
- Uncertainty propagation, i.e., allowing the inferential uncertainty in one modeling quantity to inform the inferential uncertainty in another, is a key advantage of Bayesian statistics. However, the classical tool for Bayesian estimation, Markov Chain Monte Carlo (MCMC), requires evaluating a statistical model many times, and so can be computationally expensive.

These three central data science problems may seem superfically distinct. But they share the common property that they are computationally demanding due to requiring the evaluation or estimation of a statistical model multiple times: once for each cross validation sample, once for each prior specification, or once for each draw of an MCMC chain.

I show that this commonality implies that all these tasks are amendable to *sensitivity analysis*, in which the evaluation of a model at alternative inputs is approximated by a Taylor series. By evaluating the derivatives necessary to form the Taylor series at a *single model estimate*, I avoid the re-estimation or re-evaluation that makes the above procedures computationally prohibitive. In exchange, evaluating the necessary derivatives often requires solving a large but sparse linear system, a tradeoff that can be quite favorable in practice.

Sensitivity analysis is a venerable topic, though the breadth of its applications to conteporary problems is arguably underappreciated. My work advances existing work by providing practical implementations of classical methods, particularly using automatic differentiation [Baydin et al., 2017], by updating classical theory to apply in finite sample and under more realistic conditions, and by drawing connections between superifically disparate applications of sensitivity analysis. For the remainder of the essay, I will discuss in more detail how I apply sensitivity analysis to the above three data science tasks and more, both in practice and and in theory.

Approximate cross validation. The error or variability of machine learning algorithms is often assessed by repeatedly re-fitting a model with different weighted versions of the observed data; cross-validation (CV) and the bootstrap can be thought of as examples of this technique.

In Giordano et al. [2019b], I use a linear approximation to the dependence of the fitting procedure on the weights, producing results that can be faster than repeated re-fitting by an order of magnitude. I provide explicit finite-sample error bounds for the approximation in terms of a small number of simple, verifiable assumptions. My results apply whether the weights and data are stochastic or deterministic, and so can be used as a tool for proving the accuracy of the infinitesimal jackknife on a wide variety of problems. As a corollary, I state mild regularity conditions under which our approximation consistently estimates true leave-k-out cross-validation for any fixed k. I demonstrate the accuracy of the approximation on a range of simulated and real datasets, including an unsupervised clustering problem from genomics [Luan and Li, 2003, Shoemaker et al., 2015].

Prior sensitivity for Markov Chain Monte Carlo. MCMC is arguably the most commonly used computational tool to estimate Bayesian posteriors, which is made still easier by modern black-box MCMC tools such as `Stan` [Carpenter et al., 2017, Stan Development Team, 2020]. However, a single run of MCMC typically remains time-consuming, and systematically exploring alternative prior parameterizations by re-running MCMC would be computationally prohibitive for all but the simplest models.

My software package, `rstansensitivity`, [Giordano, 2020a, Giordano et al., 2018b], takes advantage of the automatic differentiation capacities of `Stan` [Carpenter et al., 2015] together with a classical result from Bayesian robustness [Gustafson, 1996, Basu et al., 1996, Giordano et al., 2018a] to provide automatic hyperparameter sensitivity for generic `Stan` models from only a single MCMC run. I demonstrate the speed and utility of the package in detecting excess prior sensitivity, particularly in a social sciences model taken from Gelman and Hill [2006, Chapter 13.5].

Prior sensitivity for discrete Bayesian nonparameterics. A central question in many probabilistic clustering problems is how many distinct clusters are present in a particular dataset. A Bayesian nonparametric (BNP) model addresses this question by placing a generative process on cluster assignment, making the number of distinct clusters present amenable to Bayesian inference. However, like all Bayesian approaches, BNP requires the specification of a prior, and this prior may favor a greater or lesser number of distinct clusters.

In [Giordano et al., 2018c], I derive prior sensitivity measures for a truncated variational Bayes approximation using ideas from [Gustafson, 1996, Giordano et al., 2018a]. Unlike previous work on local Bayesian sensitivity for BNP [Basu, 2000], we pay special attention to the ability of our sensitivity measures to *extrapolate* to different priors, rather than treating the sensitivity as a measure of robustness *per se*. In work currently in progress, my co-author and I apply the approximation from [Giordano et al., 2018c] to an unsupervised clustering problem on a human genome dataset [Huang et al., 2011, Raj et al., 2014], demonstrating that the approximate is accurate, orders of magnitude faster than re-fitting, and capable of detecting meaningful prior sensitivity.

Uncertainty propagation in mean-field variational Bayes. Mean-field Variational Bayes (MFVB) is an approximate Bayesian posterior inference technique that is increasingly popular due to its fast runtimes on large-scale scientific data sets (e.g., Raj et al. [2014], Kucukelbir et al. [2017], Regier et al. [2019]). However, even when MFVB provides accurate posterior means for certain parameters, it often mis-estimates variances and covariances [Wang and Titterington, 2004, Turner and Sahani, 2011] due to its inability to propagate Bayesian uncertainty between statistical parameters.

In Giordano et al. [2015, 2018a], I derive a simple formula for the effect of infinitesimal model perturbations on MFVB posterior means, thus providing improved covariance estimates and greatly expanding the practical usefulness of MFVB posterior approximations. The estimates for MFVB posterior covariances rely on a result from the classical Bayesian robustness literature that relates derivatives of posterior expectations to posterior covariances and includes the Laplace approximation as a special case. In our experiments, we demonstrate that our methods are simple, general, and fast, providing accurate posterior uncertainty estimates and robustness measures with runtimes that can be an order of magnitude faster than MCMC, including models from ecology [Kéry and Schaub, 2011], the social sciences [Gelman and Hill, 2006], and on a massive internet advertising dataset [Criteo Labs, 2014].

Data ablation. In Giordano et al. [2020], I propose a method to assess the sensitivity of statistical analyses to the removal of a small fraction of the sample. Analyzing all possible data subsets of a certain size is computationally prohibitive, so I provide a finite-sample metric to approximately compute the number (or fraction) of observations that has the greatest influence on a given result when dropped. I provide explicit finite-sample error bounds on our approximation for linear and instrumental variables regressions. At minimal computational cost, the metric provides an exact finite-sample lower bound on sensitivity for any estimator, so any non-robustness our metric finds is conclusive. I demonstrate that non-robustness to data ablation is driven by a low signal-to-noise ratio in the inference problem, is not reflected in standard errors, does not disappear asymptotically, and is not a product of misspecification.

The approximation is automatically computable and works for common estimators (including OLS, IV, GMM, MLE, and variational Bayes), and I provide an easy-to-use `R` package to compute the approximation [Giordano, 2020b]. Several empirical applications based on published econometric analyses [Angelucci and De Giorgi, 2009, Finkelstein et al., 2012, Meager, 2019] show that even 2-parameter linear regression analyses of randomized trials can be highly sensitive. While I find some applications are robust, in others the sign of a treatment effect can be changed by dropping less than 1% of the sample even when standard errors are small.

Sensitivity to removal of a small fraction of the data. Classical frequentist standard errors estimate the variability in an estimator that would result from the rarefied thought experiment of re-sampling datasets from the same distribution that gave rise to the observed data. In the social sciences, this rarefied experiment rarely closely corresponds to reality, and one might be concerned if substantive conclusions could be overturned by other minor perturbations to the data.

In Giordano et al. [2020], we provide an easily-computed approximation to quantify the effect of ablating a small proportion of a dataset, with open-source software and finite-sample accuracy bounds for ordinary least squares and instrumental variables regression. We find that problems with small signal-to-noise ratio but large datasets will be particularly non-robust to the removal of a small proportion of the data. Such a situation that obtains commonly in econometrics, and we find that the sign and statistical significance of estimated effects in a number of large, prominent econometric studies can be overturned by dropping only a small number of datapoints [Angelucci and De Giorgi, 2009, Finkelstein et al., 2012, Meager, 2019].

Frequentist variability of Bayesian posteriors. Bayesian statistics provides powerful tools for coherently treating uncertainty in complex problems, though, when the model is misspecified, the estimated posterior uncertainty may not be meaningful. In principle, however, one might always compute the frequentist sampling variability of a Bayesian posterior quantity, and such a quantity always remains meaningful, even if conceptually distinct from a posterior uncertainty [Waddell et al., 2002, Kleijn and van der Vaart, 2006]. However, standard tools for evaluating frequentist uncertainty, such as the bootstrap [Huggins and Miller, 2019], are extremely computationally intensive, as they typically require re-running an MCMC procedure hundreds of times.

In a work in progress [Giordano and Broderick, 2020], we derive the Bayesian infinitesimal jackknife (IJ), which we prove can be used to consistently estimate the frequentist variability of Bayesian posterior means without bootstrapping or computing a maximum a-posteriori (MAP) estimate. Our work synthesizes results from Bayesian robustness and frequentist von Mises expansions and extends the Bayesian central limit theorem to the expectation of data-dependent functions [Jaeckel, 1972, Shao and Tu, 2012, Giordano et al., 2019b, Gustafson, 2000, Giordano et al., 2018a, Lehmann and Casella, 2006, Kass et al., 1990]. We demonstrate the accuracy of the Bayesian IJ on datasets from election modeling [Gelman and Heidemanns, 2020], ecology [Kéry and Schaub, 2011], and most of the models from [Gelman and Hill, 2006, Stan Team, 2017], showing that the Bayesian IJ can reproduce the bootstrap covariance estimates in orders of magnitude less compute time.

Sensitivity for Bayesian analysis

Selected Future work

My research is driven by the needs of my scientific collaborators, and so my future work will be determined to a large part by my colleagues. Here, I will discuss a few directions that I find promising and interesting, and which I believe could be applicable to a diverse set of problems.

The higher-order infinitesimal jackknife for the bootstrap. In the preprint Giordano et al. [2019a], we extend Giordano et al. [2019b] to higher-order Taylor series approximations, provding a family of estimators which we collectively call the higher-order infintiesimal jackknife (HOIJ). In addition to providing higher-quality approximations to CV and extending our results to k-fold CV, the higher-order approach promises to provide a scalable alternative to the bootstrap, a procedure that estimates frequentist variability by repeatedly re-evaluating a model at datasets drawn with replacement from the observed data. The bootstrap is known to enjoy higher-order accuracy in certain circumstances Hall [2013], and the HOIJ can approach the bootstrap at a rate faster than the bootstrap approaches the truth. The HOIJ thus promises to make bootstrap inference available to models which are differentiable but too expensive to re-evaluate (e.g. simulation-based models [Gourieroux and Monfort, 1993]), but also to allow efficient bootstrap-after-bootstrap procedures which that are currently out of reach for all but the simplest statsitics [Efron and Tibshirani, 1994].

Partitioned Bayesian inference. The ideas of [Giordano et al., 2018a] can be naturally extended to approximately propagate uncertainty amongst separately estimated components of an inference problem. For example, astronomical catalogues are customarily produced with MFVB-like algorithms [Lang et al., 2016, Regier et al., 2019], which take inputs such as the sky background and optical point spread function as fixed inputs, though these quantities are themselves inferred with uncertainty. Viewing all the separate inference procedures as a sequential quasi-MFVB objective, one could directly apply the techniques of LRVB to propagate the uncertainty from the modeling inputs to the astronomical catalogue's uncertainty. Doing so would require the approximate solution of a very large, but very sparse, linear system, which is itself an interesting computational challenge.

References

Angelucci, M. and De Giorgi, G. (2009). Indirect effects of an aid program: how do cash transfers affect ineligibles' consumption? *American Economic Review*, 99(1):486–508.

Basu, S. (2000). Bayesian robustness and Bayesian nonparametrics. In Insua, D. R. and Ruggeri, F., editors, *Robust Bayesian Analysis*, volume 152. Springer Science & Business Media.

Basu, S., Jammalamadaka, S. R., and Liu, W. (1996). Local posterior robustness with parametric priors: Maximum and average sensitivity. In *Maximum Entropy and Bayesian Methods*, pages 97–106. Springer.

Baydin, A., Pearlmutter, B., Radul, A., and Siskind, J. (2017). Automatic differentiation in machine learning: a survey. *Journal of Machine Learning Research*, 18(153):1–153.

Carpenter, B., Gelman, A., Hoffman, M., Lee, D., Goodrich, B., Betancourt, M., Brubaker, M., Guo, J., Li, P., and Riddell, A. (2017). Stan: A probabilistic programming language. *Journal of statistical software*, 76(1).

Carpenter, B., Hoffman, M., Brubaker, M., Lee, D., Li, P., and Betancourt, M. (2015). The stan math library: Reverse-mode automatic differentiation in c++. *arXiv preprint arXiv:1509.07164.*

Criteo Labs (2014). Criteo conversion logs dataset. Downloaded on July 27th, 2017.

Efron, B. and Tibshirani, R. (1994). *An Introduction to the Bootstrap.* CRC press.

Finkelstein, A., Taubman, S., Wright, B., Bernstein, M., Gruber, J., Newhouse, J., Allen, H., Baicker, K., and Oregon Health Study Group (2012). The oregon health insurance experiment: evidence from the first year. *The Quarterly journal of economics*, 127(3):1057–1106.

Gelman, A. and Heidemanns, M. (2020). The Economist: Forecasting the us elections. Data and model accessed Oct., 2020.

Gelman, A. and Hill, J. (2006). *Data analysis using regression and multilevel/hierarchical models.* Cambridge university press.

Giordano, R. (2020a). StanSensitivity: automated hyperparameter sensitivity for Stan models.

Giordano, R. (2020b). zaminfluence. GitHub repository `https://github.com/rgiordan/zaminfluence`.

Giordano, R. and Broderick, T. (2020). *The Bayesian Infinitesimal Jackknife for Variance.*

Giordano, R., Broderick, T., and Jordan, M. (2018a). Covariances, robustness and variational Bayes. *The Journal of Machine Learning Research*, 19(1):1981–2029.

Giordano, R., Broderick, T., and Jordan, M. I. (2015). Linear response methods for accurate covariance estimates from mean field variational Bayes. In *Advances in Neural Information Processing Systems*, pages 1441–1449.

Giordano, R., Broderick, T., and Jordan, M. I. (2018b). Automatic robustness measures in stan. Presentation at Stancon 2018 https://docs.google.com/presentation/d/1bxeFy-awELpGlDXNcdJ3r-lrKAqNCZZggD8Xmz9Omg0/edit?usp=sharing.

Giordano, R., Jordan, M. I., and Broderick, T. (2019a). A higher-order Swiss army infinitesimal jackknife. *arXiv preprint arXiv:1907.12116*.

Giordano, R., Liu, R., Jordan, M. I., and Broderick, T. (2018c). Evaluating sensitivity to the stick breaking prior in Bayesian nonparametrics. *arXiv preprint arXiv:1810.06587*.

Giordano, R., Meager, R., and Broderick, T. (2020). *An Automatic Finite-Sample Robustness Metric: Can Dropping a Little Data Change Conclusions?*

Giordano, R., Stephenson, W., Liu, R., Jordan, M. I., and Broderick, T. (2019b). A Swiss army infinitesimal jackknife. In *The 22nd International Conference on Artificial Intelligence and Statistics*, pages 1139–1147. PMLR.

Gourieroux, C. and Monfort, A. (1993). Simulation-based inference: A survey with special reference to panel data models. *Journal of Econometrics*, 59(1-2):5–33.

Gustafson, P. (1996). Local sensitivity of posterior expectations. *The Annals of Statistics*, 24(1):174–195.

Gustafson, P. (2000). Local robustness in Bayesian analysis. In Insua, D. R. and Ruggeri, F., editors, *Robust Bayesian Analysis*, volume 152. Springer Science & Business Media.

Hall, P. (2013). *The Bootstrap and Edgeworth Expansion*. Springer Science & Business Media.

Huang, L., Jakobsson, M., Pemberton, T., Ibrahim, M., Nyambo, T., Omar, S., Pritchard, J., Tishkoff, S., and Rosenberg, N. (2011). Haplotype variation and genotype imputation in African populations. *Genetic epidemiology*, 35(8):766–780.

Huggins, J. and Miller, J. (2019). Using bagged posteriors for robust inference and model criticism. *arXiv preprint arXiv:1912.07104*.

Jaeckel, L. (1972). The infinitesimal jackknife, memorandum. Technical report, MM 72-1215-11, Bell Lab. Murray Hill, NJ.

Kass, R., Tierney, L., and Kadane, J. (1990). The validity of posterior expansions based on Laplace's method. *Bayesian and Likelihood Methods in Statistics and Econometrics*.

Kéry, M. and Schaub, M. (2011). *Bayesian population analysis using WinBUGS: a hierarchical perspective.* Academic Press.

Kleijn, B. and van der Vaart, A. (2006). Misspecification in infinite-dimensional Bayesian statistics. *The Annals of Statistics*, 34(2):837–877.

Kucukelbir, A., Tran, D., Ranganath, R., Gelman, A., and Blei, D. (2017). Automatic differentiation variational inference. *The Journal of Machine Learning Research*, 18(1):430–474.

Lang, D., Hogg, D., and Mykytyn, D. (2016). The Tractor: Probabilistic astronomical source detection and measurement. *ascl*, pages ascl–1604.

Lehmann, E. and Casella, G. (2006). *Theory of point estimation.* Springer Science & Business Media.

Luan, Y. and Li, H. (2003). Clustering of time-course gene expression data using a mixed-effects model with B-splines. *Bioinformatics*, 19(4):474–482.

Meager, R. (2019). Understanding the average impact of microcredit expansions: A Bayesian hierarchical analysis of seven randomized experiments. *American Economic Journal: Applied Economics*, 11(1):57–91.

Raj, A., Stephens, M., and Pritchard, J. (2014). fastSTRUCTURE: variational inference of population structure in large SNP data sets. *Genetics*, 197(2):573–589.

Regier, J., Fischer, K., Pamnany, K., Noack, A., Revels, J., Lam, M., Howard, S., Giordano, R., Schlegel, D., and McAuliffe, J. (2019). Cataloging the visible universe through Bayesian inference in Julia at petascale. *Journal of Parallel and Distributed Computing*, 127:89–104.

Shao, J. and Tu, D. (2012). *The Jackknife and Bootstrap.* Springer Series in Statistics.

Shoemaker, J. E., Fukuyama, S., Eisfeld, A. J., Zhao, D., Kawakami, E., Sakabe, S., Maemura, T., Gorai, T., Katsura, H., Muramoto, Y., Watanabe, S., Watanabe, T., Fuji, K., Matsuoka, Y., Kitano, H., and Kawaoka, Y. (2015). An ultrasensitive mechanism regulates influenza virus-induced inflammation. *PLoS Pathogens*, 11(6):1–25.

Stan Development Team (2020). RStan: the R interface to Stan. R package version 2.21.2.

Stan Team (2017). Stan example models wiki. Referenced on June 5th, 2020.

Turner, R. E. and Sahani, M. (2011). Two problems with variational expectation maximisation for time-series models. In Barber, D., Cemgil, A. T., and Chiappa, S., editors, *Bayesian Time Series Models.*

Waddell, P., Kishino, H., and Ota, R. (2002). Very fast algorithms for evaluating the stability of ml and Bayesian phylogenetic trees from sequence data. *Genome Informatics*, 13:82–92.

Wang, B. and Titterington, M. (2004). Inadequacy of interval estimates corresponding to variational Bayesian approximations. In *Workshop on Artificial Intelligence and Statistics*, pages 373–380.